

# Minimum Spanning Trees, Perfect Matchings and Cycle Covers Over Stochastic Points in Metric Spaces

Lingxiao Huang      Jian Li  
Institute for Interdisciplinary Information Sciences  
Tsinghua University, China

December 3, 2012

## Abstract

We consider the stochastic graph model where the location of each vertex is a random point in a given metric space. We study the problems of computing the expected lengths of the minimum spanning tree, the minimum perfect matching and the minimum cycle cover on such a stochastic graph and obtain an FPRAS (Fully Polynomial Randomized Approximation Scheme) for each of these problems. Our result for stochastic minimum spanning trees improves upon the previously known constant factor approximation algorithm. Our results for the stochastic minimum perfect matching and the stochastic minimum cycle cover are the first known algorithms to the best of our knowledge.

# 1 Introduction

Motivated by the uncertainty inherent in the large graph data generated nowadays by a variety of sources, we consider the following fundamental stochastic graph model. We are given a metric space  $\mathcal{P}$ . The location of each node  $v \in \mathcal{V}$  in the stochastic graph  $\mathcal{G}$  is a random point in the metric space and the probability distribution is given as the input. We assume the distributions are discrete and independent of each other. We use  $p_{vs}$  to denote the probability that the location of node  $v$  is point  $s \in \mathcal{P}$ . The model is also termed as the *locational uncertainty model* in [23]. A special case of this model where all points follow the same distribution has been studied extensively in the stochastic geometry literature (see e.g., [7, 9, 8, 24, 29]). The model is also of fundamental interests in the area of wireless networks. In many applications, we only have some prior information about the locations of the transmission nodes (e.g., some sensors that will be deployed randomly in a designated area by an aircraft). Such a stochastic wireless network can be captured precisely by this model. See e.g., a recent survey [20] and the reference therein.

We are interested in estimating the expected length of certain combinatorial objects in the stochastic graph model. We need some notations in order to define our problems formally. We use the term *nodes* (or vertices) to refer to the vertices of the graph and *points* (or locations) the points in the metric space. Denote the set of nodes as  $\mathcal{V} = \{v_1, \dots, v_n\}$  and the set of points  $\mathcal{P} = \{s_1, \dots, s_m\}$ , where  $n = |\mathcal{V}|$  and  $m = |\mathcal{P}|$ . A realization  $\mathbf{r}$  of the stochastic graph  $\mathcal{G}$  can be represented by an  $n$ -dimensional vector  $(\mathbf{r}_1, \dots, \mathbf{r}_n) \in \mathcal{P}^n$  where point  $\mathbf{r}_i$  is the location of node  $v_i$  for  $1 \leq i \leq n$ . Let  $\mathbf{R}$  denote the set of all possible realizations. Since the nodes are independent, we can see  $\mathbf{r}$  occurs with probability  $\Pr[\mathbf{r}] = \prod_{i \in [n]} p_{v_i \mathbf{r}_i}$ . In this paper, we study three classic combinatorial problems in this model: minimum spanning tree (MST), minimum length perfect matching (MPM) (assuming an even number of nodes) and minimum length cycle cover (CC). Taking the minimum spanning tree problem for example, we would like to estimate the following quantity:

$$\mathbb{E}[\text{MST}] = \sum_{\mathbf{r} \in \mathbf{R}} \Pr[\mathbf{r}] \cdot \text{MST}(\mathbf{r})$$

where  $\text{MST}(\mathbf{r})$  is the length of the minimum spanning tree spanning all points in  $\mathbf{r}$ . However, the above formula does not give us an efficient way to estimate the expectation since it involves an exponential number of terms.

In a closely related stochastic graph model, the location of a node is a fixed point, but the existence of the node is probabilistic (called the *existential uncertainty model*). Kamousi, Chan and Suri [23] initiated the study of estimating the expected length of combinatorial objects in the existential uncertainty model. They showed that computing the expected length of the nearest neighbor (NN) graph, the Gabriel graph (GG), the relative neighborhood graph (RNG), and the Delaunay triangulation (DT) can be solved exactly in polynomial time, while computing  $\mathbb{E}[\text{MST}]$  is  $\#P$ -hard and there exists a simple FPRAS for approximating  $\mathbb{E}[\text{MST}]$ . They also gave a deterministic PTAS for approximating  $\mathbb{E}[\text{MST}]$  in Euclidean plane. They also studied the problem of computing  $\mathbb{E}[\text{MST}]$  on the locational uncertainty model. They showed the problem is also  $\#P$ -hard and gave a constant factor approximation algorithm for a special case of the problem.

## 1.1 Our Contributions

We recall that a *fully polynomial randomized approximation scheme (FPRAS)* for a problem  $f$  is a randomized algorithm  $A$  that takes an input instance  $x$  a real number  $\epsilon > 0$ , returns  $A(x)$  such that  $\Pr[(1 - \epsilon)f(x) \leq A(x) \leq (1 + \epsilon)f(x)] \geq \frac{3}{4}$  and its running time is polynomially in both the size of the input  $n$  and  $1/\epsilon$ . Perhaps the simplest and the most commonly used technique for estimating the expectation of a random variable is the naive Monte Carlo method, that is to use the sample average as the estimate. However, the method is only efficient (i.e., runs in polynomial time) if the variance of the random variable is small (More precisely,

we need the ratio between the maximum possible value and the expected value is bounded by a polynomial. See Lemma 1). To circumvent the difficulty caused by the high variance, a general methodology is to decompose the expectation of the random variable into a convex combination of conditional expectations using the law of total expectation:

$$\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}[X | Y]] = \sum_y \Pr[Y = y] \mathbb{E}[X | Y = y].$$

Hopefully, the probabilities  $\Pr[Y = y]$  can be estimated (or calculated exactly) efficiently, and the random variable  $X$  conditioning on each event  $y$  has a low variance, thus we can estimate the conditional expectation efficiently as well using Monte Carlo. However, choosing the right events  $Y$  to condition on can be tricky. For example, the FPRAS developed in [23] for estimating the expected length of the minimum spanning tree in the vertex uncertainty model follows the general conditional expectation methodology. Roughly speaking, the events to condition on are of the form “Both  $s$  and  $t$  are active (present) and  $t$  is the furthest vertex from  $s$ . In fact, conditioning on such an event, it is easy to see that the length of any spanning tree is at most  $nd(s, t)$  and at least  $d(s, t)$ . Therefore, by Chernoff bound, we can show the number of samples required for obtaining an  $1 \pm \epsilon$ -estimate for the conditional expectation can be bounded by a polynomial. In fact, we also show that the same idea can be extended to give an alternative FPRAS for the minimum spanning tree in the locational uncertainty model (Appendix A). However, it is not clear how to extend this technique for the minimum perfect matching problem and the minimum cycle cover problem. In particular, the ratio between the maximum possible length of any perfect matching (and cycle cover) and the expected length can not be bounded by fixing the positions of any constant number of vertices.

Our FPRASs for all three problems considered in this paper, the minimum spanning tree, the minimum perfect matching and the minimum cycle cover, also follow the general methodology. However, the events we choose to condition on are quite different from the previous work [23] and are quite indirect, in our opinion. Our main contributions and the highlights of our techniques can be summarized as follows:

1. (Section 2) We develop a new technique to devise FPRAS for estimating the expected length of combinatorial structures in a stochastic graph. We first demonstrate an application of this technique to the minimum spanning tree problem (MST). We obtain an FPRAS for estimating  $\mathbb{E}[\text{MST}]$ , which improves upon the previously known constant factor approximation algorithm [23]. Note the problem is known to be  $\#P$ -hard [23]. Now, we give a high level sketch of our technique. We first identifies a “core” set  $H$  of points (we call  $H$  the *home*) such that with probability close to 1, all nodes realize to  $H$ . Moreover, estimating the expectation conditioning on the event that all nodes realize to  $H$  can be done using naive Monte Carlo method since we can show the ratio between  $\max \text{MST}$  and  $\mathbb{E}[\text{MST}]$  can be bounded by a polynomial. The problematic part is when some nodes realize to points outside home. Even though the probability of such events is very small, but the length of MST under such events may be considerably large, thus contributing nontrivially to  $\mathbb{E}[\text{MST}]$ . However, we can show the contribution of such events is dominated by a subset of events where only one node realizes outside home. In other words, the contribution of the events where more than one nodes are outside home is negligible and can be safely ignored. Our technique seems more flexible and extendable than the previous technique (at least to minimum perfect matching) and we view it as a key contribution in this paper.
2. (Section 3) As a more interesting application of our “home set” technique, we give the first FPRAS (to the best of our knowledge) for approximating the expected length of the minimum perfect matching (MPM) in a stochastic graph. Our algorithm is technically more involved than the one for MST. We assume that there are even number of nodes. There are two major modifications. First, the home

set  $H$  consists of several clusters of points, so that with probability close to 1, each cluster contains even number of nodes. We can also estimate the expectation conditioning on the event that all nodes realize to  $H$  using the Monte Carlo method. Second, in order to show that the contribution of the events where more than one nodes are out of home is negligible, we need several structural properties of perfect matchings and a more careful charging argument.

3. (Section 4) We show that the problem of computing the expected length of the minimum cycle cover (CC) in a stochastic graph admits an FPRAS. We allow cycles with two nodes. It is the first known algorithm for this problem to the best of our knowledge. The event we choose to condition on is of the form “Edge  $e$  is the longest edge in the nearest neighbor graph (NN)”. Even though NN can be very different from CC, we show that, interestingly, by conditioning on such events, estimating CC becomes easier in most cases. In some cases, estimating CC is still difficult, but we can show the contribution of those cases is negligible. This is done by noticing a relationship between the length of NN and that of CC. Our algorithm can be extended to handle the case where the existence of each node is uncertain and/or each cycle is required to contain at least three nodes.

In the appendix, we show that in fact the algorithm developed in [23] can be modified into an FPRAS for estimating  $\mathbb{E}[\text{MST}]$  in the locational uncertainty model (even though only a constant approximation for a special case was claimed explicitly in that paper). However, it is not clear how their approach can be extended to perfect matching and cycle cover. All of our algorithms run in polynomial times. However, we have not attempted to optimize the exact running times.

## 1.2 Related Work

Several geometric properties of a set of stochastic points have been studied extensively in the literature under the term *stochastic geometry*. For instance, Bearwood et al. [7] shows that if there are  $n$  points uniformly and independently distributed in  $[0, 1]^2$ , the minimal traveling salesman tour visiting them has an expected length  $\Omega(\sqrt{n})$ . Asymptotic results for minimum spanning trees and minimum matchings on  $n$  points uniformly distributed in unit balls are established by Bertsimas and van Ryzin [9]. Similar results can be found in e.g., [8, 24, 29]. Compared with results in stochastic geometry, we focus on efficient computation of the statistics, instead of giving explicit mathematical formulas for them.

Recently, a number of researchers have begun to explore geometric computing under uncertainty and many classical computational geometry problems have been studied in different uncertainty models. Agarwal, Cheng, Tao and Yi [2] studied the problem of indexing probabilistic points with continuous distributions for range queries on a line. Kamousi, Chan and Suri [22] studied the closest pair and (approximate) nearest neighbor problems (i.e., finding the point with the smallest expected distance from the query point) in the existential uncertainty model. Agarwal, Efrat, Sankararaman, and Zhang [3] also studied the same problem in the locational uncertainty model under Euclidean metric. The most probable  $k$ -nearest neighbor problem and its variants have attracted a lot of attentions in the database community (See e.g., [12, 10, 26]). Several other problems have also been considered recently, such as computing the expected volume of a set of probabilistic rectangles in a Euclidean space [32], skylines (Pareto curves) over probabilistic points [6, 1], and shape fitting [28]. Instead of using probability theory, an alternative model to capture the uncertain is the *robust model*, where each point is assumed to lie in some *uncertain region* and we are interested in the extreme values of the combinatorial objects. For a comprehensive treatment of this model, see Löffler’s thesis [27] and the references therein.

The *randomly weighted graph* model where the edge weights are independent nonnegative variables has also been studied extensively. Frieze [17] and Steele [30] showed that the expected value of the minimum spanning tree on such a graph with identically and independently distributed edges is  $\zeta(3)/D$  where

$\zeta(3) = \sum_{j=1}^{\infty} 1/j^3$  and  $D$  is the derivative of the distribution at 0. Alexopoulos and Jacobson [4] developed algorithms that compute the distribution of MST and the probability that a particular edge belongs to MST when edge lengths follow discrete distributions. However, the running times of their algorithms may be exponential in the worst cases. Recently, Emek, Korman and Shavitt [16] showed that computing the  $k$ th moment of a class of properties, including the diameter, radius and minimum spanning tree, admits an FPRAS for every fixed  $k$ . Our model differs from their model in that the edge lengths in our model are not independent.

The computational/algorithmic aspects of stochastic geometry have also gained a lot of attention in recent years from the area of wireless networking. In many application scenarios, it is common to assume the nodes (e.g., sensors) are deployed randomly across a certain area, thereby forming a stochastic network. It is of central importance to study various properties in this network, such as connectivity [18], transmission capacity [19]. We refer interested reader to a recent survey [20] for more references.

## 2 Minimum Spanning Trees

In this section, we assume the presence of each node is certain but its location is stochastic. We use the term *nodes* (or *vertices*) to refer to the vertices  $\mathcal{V}$  of the spanning tree and *points* (or *locations*) the points in the metric space  $\mathcal{P}$ . We have  $|\mathcal{V}| = n$  and  $|\mathcal{P}| = m$ . We first assume the distribution of the location of each node is discrete. For any node  $v \in \mathcal{V}$  and point  $s \in \mathcal{P}$ , we use the notation  $v \models s$  to denote the event that node  $v$  is present at point  $s$ . Let  $p_{vs} = \Pr[v \models s]$ , i.e., the probability that node  $v$  is present at point  $s$ . Since node  $v$  is present with certainty, we have  $\sum_{s \in \mathcal{P}} p_{vs} = 1$ . For a point  $s$ , we let  $p(s)$  to denote the expected number of nodes realized at  $s$ , i.e.,  $\sum_{v \in \mathcal{V}} p_{vs}$ . For a set  $H$  of points, let  $p(H) = \sum_{s \in H} p(s)$ , i.e., the expected number of points realized in  $H$ . For a set  $H$  of points and a set  $S$  of nodes, we use  $H \langle S \rangle$  to denote the event that all and only nodes in  $S$  are realized to some points in  $H$ . If  $S$  only contains one node, say  $v$ , we use the notation  $H \langle v \rangle$  as the shorthand for  $H \langle \{v\} \rangle$ . Let  $H \langle i \rangle$  to denote the event  $\bigvee_{S: |S|=i} H \langle S \rangle$ , i.e., the event that exactly  $i$  nodes are in  $H$ . We use  $\text{diam}(H)$ , called the diameter of  $H$ , to denote  $\max_{s,t \in H} d(s,t)$ . Let  $d(p, H)$  denote the closest distance between point  $p$  and any point in  $H$ .

### 2.1 The Naive Monte Carlo Method

Before describing our algorithm, we first consider the naive Monte Carlo strategy, which is an important building block in our later developments. In each Monte Carlo iteration, we take a sample (a realization of all nodes), compute the length of the MST on the sample. At the end, we output the average MST lengths of all samples. The number of samples required by this algorithm is suggested by the following standard Chernoff bound.

**Lemma 1** (Chernoff Bound) *Let random variables  $X_1, X_2, \dots, X_n$  be independent random variables taking on values between 0 and  $U$ . Let  $X = \sum_{i=1}^n X_i$  and  $\mu$  be the expectation of  $X$ , for any  $\epsilon > 0$ ,*

$$\Pr[X \in [(1 - \epsilon)\mu, (1 + \epsilon)\mu]] \geq 1 - 2e^{-N \frac{\mu}{U} \epsilon^2 / 4}.$$

Therefore, for any  $\epsilon > 0$ , in order to get an  $(1 \pm \epsilon)$ -approximation with probability  $1 - \frac{1}{\text{poly}(n)}$ , the number of samples needs to be  $O(\frac{U}{\mu \epsilon^2} \log\{n\})$ . If  $\frac{U}{\mu}$ , the ratio between the maximum possible length of any MST and the expected length  $\mathbb{E}[\text{MST}]$ , is bounded by  $\text{poly}(m, n, \frac{1}{\epsilon})$  we can use the above Monte Carlo method to estimate  $\mathbb{E}[\text{MST}]$  with a polynomial number of samples. Since we use this condition often, we devote a separate definition to it.

**Definition 1** We call a random variable  $X$  a nice instance if the ratio between the maximum possible value of  $X$  and the expected value  $\mathbb{E}[X]$  is bounded by  $\text{poly}(m, n, \frac{1}{\epsilon})$ .

## 2.2 Our FPRAS for MST

We first give a high level overview of our technique. Following the general conditional expectation methodology, we break  $\mathbb{E}[\text{MST}]$  into a linear sum of conditional expectations. The events we choose to condition on depends on the notion of *home*, which is a set  $H$  of points with two nice properties: (1) with probability close to 1, all vertices are realized in home  $H$ , and (2) the ratio between the diameter of  $H$  and the expected length of MST conditioning on that all nodes are at home is bounded by a polynomial. Each event is of the form  $H\langle i \rangle$  (i.e., exactly  $i$  nodes are realized in  $H$ ) for some  $i \geq 0$ . Thus, it suffices to estimate  $\Pr[H\langle i \rangle] \mathbb{E}[\text{MST} \mid H\langle i \rangle]$  for each  $i$ . However, our final estimation only consists of the first two terms:  $\Pr[H\langle n \rangle] \mathbb{E}[\text{MST} \mid H\langle n \rangle]$  and  $\Pr[H\langle n-1 \rangle] \mathbb{E}[\text{MST} \mid H\langle n-1 \rangle]$ . We can show that the contribution from the rest of terms (where more than one nodes are outside home) is negligible and can be safely ignored. To estimate the first term. We use the second property of  $H$  which guarantees that MST conditioning on  $H\langle n \rangle$  is a nice instance. The second term can be estimated similarly.

The details of our estimation algorithm are as follows. First, we find in poly-time a set  $H$  of points (see Lemma 3 below) such that the following two properties hold:

P1.  $p(H) \geq n - \frac{\epsilon}{16}$ , and

P2.  $\mathbb{E}[\text{MST} \mid H\langle n \rangle] = \Omega(\text{diam}(H) \frac{\epsilon^2}{m^2})$ .

We call  $H$  the *home* of all nodes (due to the first property). We note that  $H$  depends on the error parameter  $\epsilon$ . Let  $F = \mathcal{P} \setminus H$ . By the law of total expectation, the expected length of the minimum spanning tree can be expanded as the following:

$$\mathbb{E}[\text{MST}] = \sum_{i \geq 0} \mathbb{E}[\text{MST} \mid F\langle i \rangle] \cdot \Pr[F\langle i \rangle].$$

Interestingly, we can show that the contribution of all terms except the first two is negligible (in Lemma 5). Therefore, it suffices to focus on estimating the first two terms

$$\mathbb{E}[\text{MST} \mid H\langle n \rangle] \cdot \Pr[H\langle n \rangle] \quad \text{and} \quad \mathbb{E}[\text{MST} \mid F\langle 1 \rangle] \cdot \Pr[F\langle 1 \rangle].$$

Now, we present the details of how to get a  $1 \pm \epsilon$ -estimate for both terms with  $O(\frac{nm^2}{\epsilon^5} \ln n)$  samples.

Estimating the first term : Due to (P2), we have a nice instance and can therefore obtain a  $1 \pm \epsilon$ -estimate for  $\mathbb{E}[\text{MST} \mid H\langle n \rangle]$  using the Monte Carlo method with  $O(\frac{nm^2}{\epsilon^4} \ln n)$  samples satisfying  $H\langle n \rangle$  (by Lemma 1). To obtain samples satisfying  $H\langle n \rangle$  efficiently, we simply use the rejection sampling method, i.e., rejecting all samples for which  $H\langle n \rangle = \text{false}$ . By the first property of  $H$ , with probability close to 1, a sample satisfies  $H\langle n \rangle$ . So, the expected time to obtain an useful sample is bounded by a constant. Overall, we can obtain a  $1 \pm \epsilon$ -estimate of the first term with using  $O(\frac{nm^2}{\epsilon^4} \ln n)$  samples with high probability.

Estimating the second term : To compute the second term, we first rewrite it as follows:

$$\begin{aligned} \mathbb{E}[\text{MST} \mid F\langle 1 \rangle] \cdot \Pr[F\langle 1 \rangle] &= \sum_{v \in \mathcal{V}} \mathbb{E}[\text{MST} \mid F\langle v \rangle] \Pr[F\langle v \rangle] \\ &= \sum_{v \in \mathcal{V}} \left( \sum_{s \in F} \Pr[F\langle v \rangle \wedge v \models s] \mathbb{E}[\text{MST} \mid F\langle v \rangle, v \models s] \right) \end{aligned}$$

Fix a node  $v$ . To estimate  $\sum_{s \in F} \Pr[F\langle v \rangle \wedge v \models s] \mathbb{E}[\text{MST} \mid F\langle v \rangle, v \models s]$ , we break it into two parts:



1. We first estimate the sum  $\sum_{s: d(s, H) < \frac{n}{\epsilon} \cdot \text{diam}(H)} \Pr[F\langle v \rangle, v \models s] \mathbb{E}[\text{MST} \mid F\langle v \rangle, v \models s]$ . Let  $\text{Cl}(v)$  be the event that  $v$  is the only node that realizes to some node  $s \notin H$  and  $d(s, H) < \frac{n}{\epsilon} \cdot \text{diam}(H)$ . Notice that the sum is in fact  $\Pr[\text{Cl}(v)] \cdot \mathbb{E}[\text{MST} \mid \text{Cl}(v)]$ . We can see that  $\Pr[\text{Cl}(v)]$  can be computed exactly in linear time. Our estimate of  $\mathbb{E}[\text{MST} \mid \text{Cl}(v)]$  is the average of  $O(\frac{m^2}{\epsilon^3} \ln n)$  samples (the samples are taken under the condition  $\text{Cl}(v)$ ). We argue the quality of this estimation is good by considering the following two cases:

- (a) Assume that  $\mathbb{E}[\text{MST} \mid \text{Cl}(v)] \geq \frac{1}{2} \mathbb{E}[\text{MST} \mid H\langle n \rangle] \geq \Omega\left(\frac{\epsilon^2}{m}\right) \text{diam}(H)$ . In this case, we have a nice instance. This is because under the condition  $\text{Cl}(v)$ , the maximum possible length of any minimum spanning tree is  $O(\frac{n}{\epsilon} \text{diam}(H))$ . Hence we can use Monte Carlo to get a  $(1 \pm \epsilon)$ -approximation of  $\mathbb{E}[\text{MST} \mid \text{Cl}(v)]$  with  $O(\frac{m^2}{\epsilon^3} \ln n)$  samples.
- (b) Otherwise, we assume that  $\mathbb{E}[\text{MST} \mid \text{Cl}(v)] \leq \frac{1}{2} \mathbb{E}[\text{MST} \mid H\langle n \rangle]$ . The probability that the sample average is larger than  $\mathbb{E}[\text{MST} \mid H\langle n \rangle]$  is at most  $\text{poly}(\frac{1}{n})$  by Chernoff Bound. The probability that for all nodes  $v$ , the sample average are at most  $\mathbb{E}[\text{MST} \mid H\langle n \rangle]$  is at least  $1 - \text{poly}(\frac{1}{n})$  by union bound. If this is the case, we can see their total contribution to the final estimation of  $\mathbb{E}[\text{MST}]$  is less than  $\epsilon \mathbb{E}[\text{MST} \mid H\langle n \rangle] \Pr[H\langle n \rangle]$ . In fact, this is because

$$\sum_{v \in V} \Pr[\text{Cl}(v)] \cdot \mathbb{E}[\text{MST} \mid \text{Cl}(v)] \leq \sum_{v \in V} \Pr[\text{Cl}(v)] \cdot \mathbb{E}[\text{MST} \mid H\langle n \rangle] < \epsilon \mathbb{E}[\text{MST} \mid H\langle n \rangle] \Pr[H\langle n \rangle].$$

The first inequality is due to the fact that  $\sum_{v \in V} \Pr[\text{Cl}(v)] \leq n - p(H) < \epsilon/16 < \epsilon \Pr[H\langle n \rangle]$ .

2. In the other part, each term has  $d(s, H) > \frac{n}{\epsilon} \cdot \text{diam}(H)$ . We just use  $d(s, H)$  as the estimation of  $\mathbb{E}[\text{MST} \mid F\langle v \rangle, v \models s]$ . This is because the length of MST is always at least  $d(s, H)$  and at most  $d(s, H) + n \cdot \text{diam}(H) \leq (1 + \epsilon)d(s, H)$ .

## 2.3 Finding Home

The remaining task is to show how to find the home  $H$ . We need the following simple lemma.

**Lemma 2** Consider two points  $s$  and  $t$  in  $\mathcal{P}$ . Suppose no node contributes to more than one half of both  $p(s)$  and  $p(t)$  (i.e.,  $\nexists v \in V$ , s.t.  $p_{vs} \geq 0.5p(s)$  and  $p_{vt} \geq 0.5p(t)$ ). Then, we have that

$$\Pr[\exists v \neq u, v \models s, u \models t] = \Omega(p(s)p(t)).$$

*Proof:* According to the given conditions, we have that

$$\frac{p_{vs}p_{vt}}{p(s)p(t)} \leq \frac{1}{4} \left( \frac{p_{vs}}{p(s)} + \frac{p_{vt}}{p(t)} \right)^2 \leq \frac{3}{8} \left( \frac{p_{vs}}{p(s)} + \frac{p_{vt}}{p(t)} \right)$$

Then, we can see that

$$\Pr[\exists v \neq u, v \models s, u \models t] = p(s)p(t) - \sum_{v \in V} p_{vs}p_{vt} \geq p(s)p(t) - \sum_{v \in V} \frac{3}{8} p(s)p(t) \left( \frac{p_{vs}}{p(s)} + \frac{p_{vt}}{p(t)} \right) \leq \frac{1}{4} p(s)p(t)$$

The last inequality holds since both  $\frac{p_{vs}}{p(s)}$  and  $\frac{p_{vt}}{p(t)}$  are at most  $1/2$ . □

With this lemma at hand, finding the home  $H$  is not difficult, as shown in the next lemma.

**Lemma 3** *There is a set  $H$  of points such that*

P1.  $p(H) \geq n - \frac{\epsilon}{16} = n - O(\epsilon)$ , and

P2.  $\mathbb{E}[\text{MST} \mid H\langle n \rangle] = \Omega\left(\text{diam}(H) \frac{\epsilon^2}{m^2}\right)$ .

*Furthermore, we can find such  $H$  in linear time.*

*Proof:* For each ordered pair of points  $(s, t)$ , consider  $H_{st} = B(s, d(s, t))$ , the ball centered at  $s$  with radius  $d(s, t)$ . Consider the furthest two points among all points  $r$  with  $p(r) \geq \frac{\epsilon}{16m}$ . Suppose the two points are  $s$  and  $t$ . For each point  $r$  that is not in  $H_{st}$ , we know  $p(r) < \frac{\epsilon}{16m}$ . Therefore, we have that  $p(\mathcal{P} \setminus H_{st}) < \frac{\epsilon}{16}$ . and  $p(H_{st}) \geq n - \frac{\epsilon}{16}$ . Consider two cases:

1. There is no node  $v \in V$  such that  $p_{vs} \geq 0.5p(s)$  and  $p_{vt} \geq 0.5p(t)$ . In this case, by Lemma 2, we have that

$$\mathbb{E}[\text{MST} \mid H_{st}\langle n \rangle] \geq d(s, t) \Pr[\exists v \neq u, v \models s, u \models t] \geq \Omega\left(d(s, t) \frac{\epsilon^2}{m^2}\right).$$

2. There is a node  $v$  such that  $p_{vs} \geq 0.5p(s)$  and  $p_{vt} \geq 0.5p(t)$ . In this case, conditioning on the event that a different node  $u$  is realized to an arbitrary point  $q$

$$\mathbb{E}[\text{MST} \mid H_{st}\langle n \rangle] \geq d(s, q) \Pr[v \models s] + d(t, q) \Pr[v \models t] \geq d(s, t) \frac{\epsilon}{32m}.$$

In either case,  $H_{st}$  satisfies both P1 and P2. □

## 2.4 Analysis of the Performance Guarantee

Now, we analyze the performance guarantee of our algorithm. We need to show that the total contribution from the scenarios where more than one nodes are not at home is very small. We need some notations first. Suppose  $S$  is the set of nodes out of home  $H$ . We use  $\mathcal{F}_S$  to denote the set of all possible realizations of all nodes in  $S$  to points in  $F$  (we can think each element in  $\mathcal{F}_S$  as a  $|S|$ -dimensional vector where each coordinate is indexed by a vertex in  $S$  and its value is a point in  $F$ ). Similarly, we denote the set of realizations of  $\bar{S} = V \setminus S$  to points in  $H$  by  $\mathcal{H}_{\bar{S}}$ . For any  $F_S \in \mathcal{F}_S$  and  $H_{\bar{S}} \in \mathcal{H}_{\bar{S}}$ , we use  $(F_S, H_{\bar{S}})$  to denote the event that both  $F_S$  and  $H_{\bar{S}}$  happen and  $\text{MST}(F_S, H_{\bar{S}})$  the length of the minimum spanning tree under the realization  $(F_S, H_{\bar{S}})$ . We need the following combinatorial fact.

**Lemma 4** *Consider a particular realization  $(F_S, H_{\bar{S}})$  where  $S$  is the set of nodes out of home  $H$ .  $|S| \geq 2$ . The realization  $(F_{S'}, H_{\bar{S}'})$  is obtained from  $(F_S, H_{\bar{S}})$  by sending home the node that is outside  $H$  but closest to any node in  $H_{\bar{S}}$ . Then  $\text{MST}(F_S, H_{\bar{S}}) \leq 4\text{MST}(F_{S'}, H_{\bar{S}'})$ .*

*Proof:* For  $(F_S, H_{\bar{S}})$ , Let  $d = \min_{v \in F_S, u \in H_{\bar{S}}} \{d(u, v)\}$ . Then we have

$$2\text{MST}(F_{S'}, H_{\bar{S}'}) \geq \text{MST}(F_{S'}, H_{\bar{S}'}) + d \geq \frac{1}{2}\text{MST}(F_{S'}, H_{\bar{S}}) + d \geq \frac{1}{2}\text{MST}(F_S, H_{\bar{S}})$$

The second inequality holds since the length of the minimum spanning tree is at most two times the length of the minimum Steiner tree (We can think  $\text{MST}(F_{S'}, H_{\bar{S}'})$  as a Steiner tree connecting all nodes in  $F_{S'} \cup H_{\bar{S}'}$ ). □

The following lemma is essential in establishing the performance guarantee.



**Lemma 5** For any  $\epsilon > 0$ , if  $H$  satisfies the properties in Lemma 3, we have that

$$\sum_{i>1} \mathbb{E}[\text{MST} \mid F\langle i \rangle] \cdot \Pr[F\langle i \rangle] \leq \epsilon \cdot \mathbb{E}[\text{MST} \mid F\langle 1 \rangle] \cdot \Pr[F\langle 1 \rangle].$$

*Proof:* We claim that for any  $i > 1$ ,

$$\mathbb{E}[\text{MST} \mid F\langle i+1 \rangle] \cdot \Pr[F\langle i+1 \rangle] \leq \frac{\epsilon}{2} \mathbb{E}[\text{MST} \mid F\langle i \rangle] \cdot \Pr[F\langle i \rangle].$$

If the claim is true, then we can show the lemma easily by noticing that, for any  $n \geq 2$ ,

$$\sum_{i>1} \mathbb{E}[\text{MST} \mid F\langle i \rangle] \Pr[F\langle i \rangle] \leq \sum_{i=1}^{n-1} \left(\frac{\epsilon}{2}\right)^i \mathbb{E}[\text{MST} \mid F\langle 1 \rangle] \Pr[F\langle 1 \rangle] \leq \epsilon \mathbb{E}[\text{MST} \mid F\langle 1 \rangle] \Pr[F\langle 1 \rangle].$$

First, we rewrite the LHS as

$$\mathbb{E}[\text{MST} \mid F\langle i+1 \rangle] \cdot \Pr[F\langle i+1 \rangle] = \sum_{|S|=i+1} \sum_{F_S \in \mathcal{F}_S} \sum_{H_{\bar{S}} \in \mathcal{H}_{\bar{S}}} (\Pr[(F_S, H_{\bar{S}})] \cdot \text{MST}(F_S, H_{\bar{S}})).$$

Similarly, we have the RHS written as

$$\mathbb{E}[\text{MST} \mid F\langle i \rangle] \cdot \Pr[F\langle i \rangle] = \sum_{|S'|=i} \sum_{F_{S'} \in \mathcal{F}_{S'}} \sum_{H_{\bar{S}'} \in \mathcal{H}_{\bar{S}'}} (\Pr[(F_{S'}, H_{\bar{S}'})] \cdot \text{MST}(F_{S'}, H_{\bar{S}'})).$$

For each pair  $(F_S, H_{\bar{S}})$ , let  $C(F_S, H_{\bar{S}}) = \Pr[S \models F_S \wedge \bar{S} \models H_{\bar{S}}] \cdot \text{MST}(F_S, H_{\bar{S}})$ . Think each pair  $(F_S, H_{\bar{S}})$  with  $|S| = i+1$  as a seller and each pair  $(F_{S'}, H_{\bar{S}'})$  with  $|S'| = i$  as a buyer. The seller  $(F_S, H_{\bar{S}})$  want to sell the term  $C(F_S, H_{\bar{S}})$  and the buyers want to buy all these terms. The buyer  $(F_{S'}, H_{\bar{S}'})$  has a budget of  $C(F_{S'}, H_{\bar{S}'})$ . We show there is a charging scheme such that every term  $C(F_S, H_{\bar{S}})$  is fully paid by the buyers and each buyer spends at most an  $\frac{\epsilon}{n}$  fraction of her budget. Note that the existence of such a charging scheme suffices to prove the lemma.

Suppose we are selling the term  $C(F_S, H_{\bar{S}})$ . Consider the following charging scheme. Suppose  $v \in S$  is the node closest to any node in  $H_{\bar{S}}$ . Let  $S' = S \setminus v$  and  $F_{S'}$  be the restriction of  $F_S$  to all coordinates in  $S$  except  $v$ . We say  $(F_{S'}, H_{\bar{S}'})$  is consistent with  $(F_S, H_{\bar{S}})$ , denoted as  $(F_{S'}, H_{\bar{S}'} ) \sim (F_S, H_{\bar{S}})$ , if  $H_{\bar{S}}$  agrees with  $H_{\bar{S}'}$  for all vertices in  $\bar{S}$ . and  $F_S$  agrees with  $F_{S'}$  for all vertices in  $S'$ . Intuitively,  $(F_{S'}, H_{\bar{S}'})$  can be obtained from  $(F_S, H_{\bar{S}})$  by sending  $v$  to an arbitrary point in the home. Let

$$Z(F_S, H_{\bar{S}}) = \sum_{(F_{S'}, H_{\bar{S}'}) \sim (F_S, H_{\bar{S}})} \Pr[(F_{S'}, H_{\bar{S}'})].$$

For each buyer  $(F_{S'}, H_{\bar{S}'}) \sim (F_S, H_{\bar{S}})$ , we charge her the following amount of money

$$\frac{\Pr[(F_{S'}, H_{\bar{S}'})]}{Z(F_S, H_{\bar{S}})} C(F_S, H_{\bar{S}})$$

It is easy to see that  $C(F_S, H_{\bar{S}})$  is fully paid by all buyers consistent with  $(F_S, H_{\bar{S}})$ . It remains to show that each buyer  $(F_{S'}, H_{\bar{S}'})$  has been charged at most  $\frac{\epsilon}{n} C(F_{S'}, H_{\bar{S}'})$ . By the above charging scheme, the terms

in LHS that are charged to buyer  $(F_{S'}, H_{\bar{S}'})$  are consistent with  $(F_{S'}, H_{\bar{S}'})$ . Therefore, the total amount charged to buyer  $(F_{S'}, H_{\bar{S}'})$  is

$$\begin{aligned}
& \sum_{(F_S, H_{\bar{S}}) \sim (F_{S'}, H_{\bar{S}'})} \frac{\Pr[F_{S'}, H_{\bar{S}'}]}{Z(F_S, H_{\bar{S}})} C(F_S, H_{\bar{S}}) \\
& \leq 4\text{MST}(F_{S'}, H_{\bar{S}'}) \cdot \sum_{(F_S, H_{\bar{S}}) \sim (F_{S'}, H_{\bar{S}'})} \frac{\Pr[F_{S'}, H_{\bar{S}'}]}{Z(F_S, H_{\bar{S}})} \Pr[(F_S, H_{\bar{S}})] \\
& = 4\text{MST}(F_{S'}, H_{\bar{S}'}) \Pr[F_{S'}, H_{\bar{S}'}] \cdot \sum_{(F_S, H_{\bar{S}}) \sim (F_{S'}, H_{\bar{S}'})} \frac{\Pr[F_S, H_{\bar{S}}]}{Z(F_S, H_{\bar{S}})} \\
& \leq 4\text{MST}(F_{S'}, H_{\bar{S}'}) \Pr[F_{S'}, H_{\bar{S}'}] \cdot \sum_{v \in \bar{S}'} \frac{\Pr(v \in F)}{\Pr(v \in H)} \\
& \leq \frac{\epsilon}{2} \text{MST}(F_{S'}, H_{\bar{S}'}) \Pr[F_{S'}, H_{\bar{S}'}]
\end{aligned}$$

The first inequality follows from Lemma 4. To see the second inequality, for a fixed vertex  $v$ , consider the quantity  $\sum_{(F_S, H_{\bar{S}}) \sim (F_{S'}, H_{\bar{S}'})} \frac{\Pr[F_S, H_{\bar{S}}]}{Z(F_S, H_{\bar{S}})}$ . By the definition of  $Z$ , we can see that the denominators of all terms are in fact the same. Canceling out the same multiplicative terms from the numerators and the denominator, we can see it is at most  $\frac{\Pr(v \in F)}{\Pr(v \in H)}$ .  $\square$

In sum, we obtain the following theorem.

**Theorem 1** *There is an FPRAS for estimating the expected length of the minimum spanning tree in a stochastic graph.*

Finally, we note that we can use the  $O(n \text{poly}(1/\epsilon) \text{poly} \log n)$  time algorithm to estimate the  $1 + \epsilon$ -approximate value of the minimum spanning tree [15], instead of the exact algorithms, in a general metric graph. For Euclidean spaces with  $O(1)$  dimensions, we can use the  $O(\sqrt{n} \text{poly}(1/\epsilon) \text{poly} \log n)$  time algorithm in [14] for computing such an approximate value (under certain assumptions) or the  $O(n \text{poly}(1/\epsilon))$  time algorithm in [11].

### 3 Minimum Perfect Matchings

In this section, we consider the minimum perfect matching problem. We assume the number of nodes,  $n$ , is even. For a node  $v$  and a set  $H$  of points, let  $p_v(H) = \sum_{s \in H} p_{vs}$ . For two sets  $H_1$  and  $H_2$  of points, let  $d(H_1, H_2) = \min_{s \in H_1, t \in H_2} \{d(s, t)\}$ . We use MPM to denote the length of the minimum length perfect matching. Our goal is to estimate  $\mathbb{E}[\text{MPM}]$ .

#### 3.1 Our FPRAS for MPM

Our algorithm for MPM follows the same framework: We first identify the home such that the conditional expectation conditioning on all nodes are at home can be estimated using the Monte Carlo method. We can similarly show that the contribution from the scenarios where more than one nodes are outside home is negligible. Thus, we only need to estimate two parts: (1) the expectation conditioning on that all nodes are at home, and (2) the expectation conditioning on that only one node is not at home. There are two major differences from the algorithm for MST. First, the home set is composed by several clusters of points, instead of a single ball. Second, we need a more careful charging argument.

Now, we present the details of our algorithm. First, we find a collection of sets of points  $H_1, \dots, H_k$  such that the following properties holds.

- Q1. For each node  $v$ , there is a (unique) set  $H_j$  such that  $p_v(H_j) \geq 1 - O(\frac{\epsilon}{nm^3})$ . We call  $H_j$  the *home* of node  $v$ , denoted as  $H(v)$ .
- Q2. For each ball  $H_j$ , the number of nodes  $v$  with  $H_j$  as its home (i.e.,  $H(v) = H_j$ ) is even.
- Q3.  $\mathbb{E}[\text{MPM}] = \Omega(\frac{\epsilon D}{nm^5})$  where  $D = \max_i \{\text{diam}(H_i)\}$ .

Let  $H = \cup_i H_i$  and  $F = \mathcal{P} \setminus H$ . We use  $H\langle n \rangle$  to denote the event for all  $n$  nodes  $v$ ,  $v \models H(v)$ . We denote the event that there are exactly  $i$  nodes which realize out of their homes by  $F\langle i \rangle$ . By the previous discuss, we only focus on estimating two terms:  $\mathbb{E}[\text{MPM} \mid H\langle n \rangle] \cdot \Pr[H\langle n \rangle]$  and  $\mathbb{E}[\text{MPM} \mid F\langle 1 \rangle] \cdot \Pr[F\langle 1 \rangle]$ .

Estimating the first term : Note that  $\Pr[H\langle n \rangle]$  is close to 1 (by union bound) and can be computed exactly. To estimate  $\mathbb{E}[\text{MPM} \mid H\langle n \rangle]$ , we take the average of  $O(\frac{n^2 m^5}{\epsilon^4} \ln n)$  samples. We distinguish the following two cases.

1.  $\mathbb{E}[\text{MPM} \mid H\langle n \rangle] \geq \frac{\epsilon}{2} \mathbb{E}[\text{MPM}] \geq \Omega(\frac{\epsilon^2 D}{nm^5})$ . We could get a  $(1 \pm \epsilon)$ -approximation using the Monte Carlo method with  $O(\frac{n^2 m^5}{\epsilon^4} \ln n)$  samples. This is because the maximum possible MPM length is at most  $nD$  and therefore we have a nice instance.
2.  $\mathbb{E}[\text{MPM} \mid H\langle n \rangle] < \frac{\epsilon}{2} \mathbb{E}[\text{MPM}]$ . Then the probability that the sample average is larger than  $\epsilon \mathbb{E}[\text{MPM}]$  is at most  $\text{poly}(\frac{1}{n})$  by Chernoff Bound. We can thus ignore this part safely.

Estimating the second term : We rewrite the second term as follows:

$$\mathbb{E}[\text{MPM} \mid F\langle 1 \rangle] \cdot \Pr[F\langle 1 \rangle] = \sum_{v \in \mathcal{V}} \left( \sum_{s \notin H(v)} \Pr[F\langle v \rangle \wedge v \models s] \mathbb{E}[\text{MPM} \mid F\langle v \rangle, v \models s] \right)$$

Fix a particular node  $v$ . We break the sum into two parts as in the previous section:

$\sum_{s: d(s, H(v)) < \frac{n}{\epsilon} \cdot \text{diam}(H)} \Pr[F\langle v \rangle, v \models s] \mathbb{E}[\text{MPM} \mid F\langle v \rangle, v \models s]$  and  $\sum_{s: d(s, H(v)) \geq \frac{n}{\epsilon} \cdot \text{diam}(H)} \Pr[F\langle v \rangle, v \models s] \mathbb{E}[\text{MPM} \mid F\langle v \rangle, v \models s]$ . For the first part, we use Monte Carlo and for the second part, we use  $d(s, H(v))$  as the estimate of  $\mathbb{E}[\text{MPM} \mid F\langle v \rangle, v \models s]$ . The details are exactly the same as in the previous section and omitted here.

### 3.2 Finding Homes

What remains now is to show how to find the home sets  $H_1, \dots, H_k$  in poly-time. We need the following lemma which is useful in bounding  $\mathbb{E}[\text{MPM}]$  from below.

**Lemma 6** *For any two disjoint sets  $H_1$  and  $H_2$  of points, and any node  $v$ , we have*

$$\mathbb{E}[\text{MPM}] \geq \frac{\min\{p_v(H_1), p_v(H_2)\}}{m} \cdot d(H_1, H_2).$$

*Proof:* Suppose  $s = \arg \max_{s'} \{p_{vs'} \mid s' \in H_1\}$ , and  $t = \arg \max_{t'} \{p_{vt'} \mid t' \in H_2\}$ . Obviously, we have  $p_{vs} \geq \frac{p_v(H_1)}{m}$  and  $p_{vt} \geq \frac{p_v(H_2)}{m}$ . So it suffices to show  $\mathbb{E}[\text{MPM}] \geq \min\{p_{vs}, p_{vt}\} \cdot d(s, t)$ . We first see that

$$\begin{aligned} \mathbb{E}[\text{MPM}] &\geq p_{vs} \mathbb{E}[\text{MPM} \mid v \models s] + p_{vt} \mathbb{E}[\text{MPM} \mid v \models t] \\ &\geq \min\{p_{vs}, p_{vt}\} \left( \mathbb{E}[\text{MPM} \mid v \models s] + \mathbb{E}[\text{MPM} \mid v \models t] \right). \end{aligned}$$

Then it is sufficient to prove that  $\mathbb{E}[\text{MPM} \mid v \models s] + \mathbb{E}[\text{MPM} \mid v \models t] \geq d(s, t)$ . Fix a realization of all nodes except  $v$  and condition on this event. Consider the two minimum perfect matchings, one for the case  $v \models s$ , (denoted as  $\text{MPM}_1$ ) and the other one for  $v \models t$  (denoted as  $\text{MPM}_2$ ). Consider the symmetric difference

$$\text{MPM}_1 \oplus \text{MPM}_2.$$

We can see that it is a path  $(s, p_1, p_2, \dots, p_k, t)$ , such that  $(s, p_1) \in \text{MPM}_1, (p_1, p_2) \in \text{MPM}_2, \dots, (p_k, t) \in \text{MPM}_2$ . So  $\text{MPM}_1 + \text{MPM}_2 \geq d(s, t)$  by the triangle inequality. Therefore, we have  $\mathbb{E}[\text{MPM} \mid v \models s] + \mathbb{E}[\text{MPM} \mid v \models t] \geq d(s, t)$ .  $\square$

Now, we are ready to show how to find the home sets in polynomial time.

**Lemma 7** *We can find in poly-time disjoint point sets  $H_1, \dots, H_k$  such that*

- Q1. For each node  $v$ , there is a unique ball  $H_j$  such that  $p_v(H_j) \geq 1 - O(\frac{\epsilon}{nm^3})$ ;*
- Q2. For all  $j$ ,  $|\{v \in \mathcal{V} \mid H(v) = H_j\}|$  is even; and*
- Q3.  $\mathbb{E}[\text{MPM}] = \Omega(\frac{\epsilon D}{nm^5})$ .*

*Proof:* We gradually increase  $t$ , starting from 0. Consider the balls  $B(s, t)$  for all points  $s$  in  $\mathcal{P}$ . Initially, each ball is a singleton component. As we increase  $t$ , if two different components intersect, we merge them into a new component. Consider the first time  $T$  such that Q1 and Q2 are satisfied by those components. Let those components be  $H_1, \dots, H_k$ . Note that such  $T$  must exist, because the set of all points satisfies the first two properties. Now, we show the Q3 also holds.

Recall  $D = \max_i \text{diam}(H_i)$ . Firstly, note that  $D \leq 2mT$ . Secondly, consider  $T' = T - \epsilon$  for some infinitesimal  $\epsilon > 0$ . At time  $T'$ , consider two situations:

1. There exists a node  $v$ , such that  $\forall j, p_v(H_j) < 1 - O(\frac{\epsilon}{nm^3})$ . Then there must exist two components  $C_1$  and  $C_2$  such that  $p_v(C_1) > \Omega(\frac{\epsilon}{nm^3})$  and  $p_v(C_2) > \Omega(\frac{\epsilon}{nm^3})$ . Moreover, since  $C_1$  and  $C_2$  are two distinct components,  $d(C_1, C_2) \geq 2T'$ . Then, by Lemma 6, we have  $\mathbb{E}[\text{MPM}] \geq \Omega(\frac{\epsilon}{nm^4}) \cdot 2T \geq \Omega(\frac{\epsilon}{nm^5})$ .
2. Suppose the Q1 is true but Q2 is still false. Suppose  $H_j$  is a component which homes odd number of nodes. We note that with probability at least  $(1 - \frac{1}{nm^3})^n \approx 1$ , every node realizes to a point in its home. When this is the case, there is at least one node in  $H_j$  that needs to be matched with some node outside  $H_j$ , which incurs a cost of at least  $2T$ .  $\square$

### 3.3 Analysis of the Performance Guarantee

We show for  $i > 1$ , the contribution from event  $F\langle i \rangle$  is negligible. We need the following structural result about minimum perfect matchings, which is essential for our charging argument.

Suppose  $S$  is the set of nodes that are out of their homes. We use  $\mathcal{F}_S$  and  $\mathcal{H}_{\bar{S}}$  to denote the set of all realizations of the all nodes in  $S$  to points in  $F$ , and the set of realizations of  $\bar{S} = V \setminus S$  to points in  $H$  respectively. We use  $\text{MPM}(\mathcal{F}_S, \mathcal{H}_{\bar{S}})$  to denote the length of the minimum perfect matching under the realization  $(\mathcal{F}_S, \mathcal{H}_{\bar{S}})$ . The following combinatorial fact plays the same role in the charging argument as Lemma 4 does in the previous section. Different from the MST problem, we can not achieve a similar bound to the one in Lemma 4 since  $\text{MPM}(\mathcal{F}_S, \mathcal{H}_{\bar{S}})$  may decrease significantly if we only sending only one node outside home back to its home. However, we show that in such case, if we send one more node back home,  $\text{MPM}(\mathcal{F}_S, \mathcal{H}_{\bar{S}})$  can still be bounded.

**Lemma 8** Fix a realization  $(F_S, H_{\bar{S}})$ . We use  $\ell(v)$  to denote  $d(v, H(v))$  for all nodes  $v \in S$ . Suppose  $v_1 \in S$  has the smallest  $\ell$  value and  $v_2$  has the second smallest  $\ell$  value. Let  $S' = S \setminus \{v_1\}$ ,  $S'' = S' \setminus \{v_2\}$ . Further let  $(F_{S'}, H_{\bar{S}'})$  be a realization obtained from  $(F_S, H_{\bar{S}})$  by sending  $v_1$  to a point in its home  $H(v_1)$  and  $(F_{S''}, H_{\bar{S}''})$  be a realization obtained from  $(F_{S'}, H_{\bar{S}'})$  by sending  $v_2$  to a point in its home  $H(v_2)$ . Then we have that

$$\text{MPM}(F_S, H_{\bar{S}}) \leq 2(m+2)\text{MPM}(F_{S'}, H_{\bar{S}'}) + 2(m+2)\text{MPM}(F_{S''}, H_{\bar{S}''})$$

*Proof:* Let  $d = \min_v \ell(v)$  and  $D = \max_i \text{diam}(H_i)$ . Note that  $d \geq \frac{D}{m}$  as  $d \geq 2T$  and  $D \leq 2mT$ . We distinguish the following three cases:

1.  $\text{MPM}(F_S, H_{\bar{S}}) \leq \frac{d}{2}$ . Using a similar argument to the one in Lemma 6, we have

$$\text{MPM}(F_{S'}, H_{\bar{S}'}) + \text{MPM}(F_S, H_{\bar{S}}) \geq \ell(v) = d$$

So, we have  $\text{MPM}(F_S, H_{\bar{S}}) \leq \text{MPM}(F_{S'}, H_{\bar{S}'})$  in this case.

2.  $\text{MPM}(F_S, H_{\bar{S}}) \geq (m+2)d$ . By the triangle inequality, we can see that

$$\text{MPM}(F_{S'}, H_{\bar{S}'}) + (m+1)d \geq \text{MPM}(F_{S'}, H_{\bar{S}'}) + d + D \geq \text{MPM}(F_S, H_{\bar{S}})$$

So, we have  $\text{MPM}(F_S, H_{\bar{S}}) \leq (m+2)\text{MPM}(F_{S'}, H_{\bar{S}'})$ .

3.  $\frac{d}{2} \leq \text{MPM}(F_S, H_{\bar{S}}) \leq (m+2)d$ .

(a)  $\text{MPM}(F_{S'}, H_{\bar{S}'}) \geq \frac{d}{2}$ . We directly have  $\text{MPM}(F_S, H_{\bar{S}}) \leq 2(m+2)\text{MPM}(F_{S'}, H_{\bar{S}'})$ .

(b)  $\text{MPM}(F_{S'}, H_{\bar{S}'}) \leq \frac{d}{2}$ . By Lemma 6, we have

$$\text{MPM}(F_{S'}, H_{\bar{S}'}) + \text{MPM}(F_{S''}, H_{\bar{S}''}) \geq d$$

Then we have  $\text{MPM}(F_S, H_{\bar{S}}) \leq 2(m+2)\text{MPM}(F_{S''}, H_{\bar{S}''})$ .

So we prove the lemma. □

What remains is to establish the following key lemma. The proof is similar to, but more involved than that of Lemma 5.

**Lemma 9** For any  $\epsilon > 0$ , if  $H$  satisfies the properties in Lemma 7, we have that

$$\sum_{i>1} \mathbb{E}[\text{MPM} \mid F\langle i \rangle] \cdot \Pr[F\langle i \rangle] \leq \epsilon \cdot \mathbb{E}[\text{MPM} \mid F\langle 0 \rangle] \cdot \Pr[F\langle 0 \rangle] + \epsilon \cdot \mathbb{E}[\text{MPM} \mid F\langle 1 \rangle] \cdot \Pr[F\langle 1 \rangle].$$

*Proof:* We claim that for any  $i > 1$ ,

$$\mathbb{E}[\text{MPM} \mid F\langle i+1 \rangle] \cdot \Pr[F\langle i+1 \rangle] \leq \frac{\epsilon}{6} (\mathbb{E}[\text{MPM} \mid F\langle i \rangle] \cdot \Pr[F\langle i \rangle] + \mathbb{E}[\text{MPM} \mid F\langle i-1 \rangle] \cdot \Pr[F\langle i-1 \rangle])$$

If the claim is true, the lemma can be proven easily as follows. For ease of notation, we use  $A(i)$  to denote  $\mathbb{E}[\text{MPM} \mid F\langle i \rangle] \cdot \Pr[F\langle i \rangle]$ . First, we can see that

$$A(i+2) + A(i+1) \leq \frac{\epsilon}{6}A(i+1) + \frac{2\epsilon}{6}A(i) + \frac{\epsilon}{6}A(i-1) \leq \frac{\epsilon}{2}(A(i) + A(i-1)).$$

So if  $i$  is odd,  $A(i+2) + A(i+1) \leq (\frac{\epsilon}{2})^{(i+1)/2}(A(1) + A(0))$ . Therefore,  $\sum_{i>1} A(i) \leq \frac{\epsilon/2}{1-\epsilon/2}(A(1) + A(0)) \leq \epsilon(A(1) + A(0))$ . Now, we prove the claim. Again, we rewrite the LHS as

$$\mathbb{E}[\text{MPM} \mid F\langle i+1 \rangle] \cdot \Pr[F\langle i+1 \rangle] = \sum_{|S|=i+1} \sum_{F_S} \sum_{H_{\bar{S}}} \left( \Pr[F_S, H_{\bar{S}}] \cdot \text{MPM}(F_S, H_{\bar{S}}) \right).$$

Similarly, we have the RHS to be

$$\mathbb{E}[\text{MPM} \mid F\langle i \rangle] \cdot \Pr[F\langle i \rangle] = \sum_{|S'|=i} \sum_{F_{S'}} \sum_{H_{\bar{S}'}} \left( \Pr[F_{S'}, H_{\bar{S}'}] \cdot \text{MPM}(F_{S'}, H_{\bar{S}'}) \right)$$

$$\mathbb{E}[\text{MPM} \mid F\langle i-1 \rangle] \cdot \Pr[F\langle i-1 \rangle] = \sum_{|S''|=i-1} \sum_{F_{S''}} \sum_{H_{\bar{S}''}} \left( \Pr[F_{S''}, H_{\bar{S}''}] \cdot \text{MPM}(F_{S''}, H_{\bar{S}''}) \right)$$

Let  $C(F_S, H_{\bar{S}}) = \Pr[F_S, H_{\bar{S}}] \cdot \text{MPM}(F_S, H_{\bar{S}})$ . Think all  $(F_{S'}, H_{\bar{S}'})$  with  $|S'| = i$  and all  $(F_{S''}, H_{\bar{S}'})$  with  $|S''| = i-1$  as buyers. The buyers want to buy all terms in LHS. The budget of buyer  $(F_{S'}, H_{\bar{S}'})$  is  $C(F_{S'}, H_{\bar{S}'})$ . We show there is a charging scheme such that every term  $C(F_S, H_{\bar{S}})$  is fully paid by the buyers and each buyer spends at most an  $\frac{\epsilon}{n}$  fraction of her budget.

Suppose we are selling the term  $C(F_S, H_{\bar{S}})$ . Consider the following charging scheme. Suppose  $v \in S$  the node that realizes to point  $f \in F_S$  which is the closest point to  $H_{\bar{S}}$  in  $F_S$ . Suppose  $t \in S$  the node that is realized to point  $F_t \in F_S$  which is the second closest point to  $H_{\bar{S}}$  in  $F_S$ . Let  $S' = S \setminus \{v_1\}$ ,  $S'' = S' \setminus \{v_2\}$ . If  $(F_{S'}, R_{\bar{S}'})$  is obtained from  $(F_S, H_{\bar{S}})$  by sending  $v_1$  to a point in its home  $H(v_1)$ , we say  $(F_{S'}, R_{\bar{S}'})$  is consistent with  $(F_S, R_{\bar{S}})$ , denoted as  $(F_{S'}, R_{\bar{S}'}) \sim (F_S, R_{\bar{S}})$ . If  $(F_{S''}, R_{\bar{S}'})$  is obtained from  $(F_{S'}, H_{\bar{S}'})$  by sending  $v_2$  to a point in its home  $H(v_2)$ , we say  $(F_{S''}, R_{\bar{S}'})$  is consistent with  $(F_{S'}, R_{\bar{S}'})$ , denoted as  $(F_{S''}, R_{\bar{S}'}) \sim (F_{S'}, R_{\bar{S}'})$ . Let

$$Z(F_S, H_{\bar{S}}) = \sum_{(F_{S'}, H_{\bar{S}'}) \sim (F_S, H_{\bar{S}})} \Pr[(F_{S'}, H_{\bar{S}'}], \quad \text{and} \quad Z(F_{S'}, H_{\bar{S}'}) = \sum_{(F_{S''}, H_{\bar{S}'}) \sim (F_{S'}, H_{\bar{S}'})} \Pr[F_{S''}, H_{\bar{S}''}]$$

For each buyer  $(F_{S'}, H_{\bar{S}'}) \sim (F_S, H_{\bar{S}})$ , we charge  $(F_{S'}, H_{\bar{S}'})$  the following amount of dollars

$$\frac{\Pr[F_{S'}, H_{\bar{S}'}]}{Z(F_S, H_{\bar{S}})} \Pr[F_S, H_{\bar{S}}] \cdot 2(m+2) \text{MPM}(F_{S'}, H_{\bar{S}'})$$

and we charge every buyer  $(F_{S''}, H_{\bar{S}'})$  consistent with  $(F_{S'}, H_{\bar{S}'})$  the following amount of money

$$\frac{\Pr[F_S, H_{\bar{S}}]}{Z(F_S, H_{\bar{S}})} \cdot \frac{\Pr[F_{S'}, H_{\bar{S}'}]}{Z(F_{S'}, H_{\bar{S}'})} \Pr[F_{S''}, H_{\bar{S}''}] \cdot 2(m+2) \text{MPM}(F_{S''}, H_{\bar{S}'})$$

In this case, we say  $(F_{S''}, H_{\bar{S}'})$  is a *sub-buyer* of the term  $C(F_S, H_{\bar{S}})$ . By Lemma 8, we can see that  $A(F_S, H_{\bar{S}})$  is fully paid. To prove the claim, it suffices to show that each buyer  $(F_{S'}, H_{\bar{S}'})$  and each sub-buyer  $(F_{S''}, H_{\bar{S}'})$  has been charged at most  $\frac{\epsilon}{n} A(F_{S'}, H_{\bar{S}'})$  dollars. By the above charging scheme, the terms in LHS that are charged to buyer  $(F_{S'}, H_{\bar{S}'})$  are consistent with  $(F_{S'}, H_{\bar{S}'})$ . Using the same argument as in the previous section, we can show the spending of  $(F_{S'}, H_{\bar{S}'})$  as a buyer is at most

$$\frac{\epsilon}{nm} \text{MPM}(F_{S'}, H_{\bar{S}'}) \Pr[F_{S'}, H_{\bar{S}'}].$$



The spending of  $(F_{S''}, H_{\bar{S}''})$  as a sub-buyer can be bounded as follows:

$$\begin{aligned}
& \sum_{(F_S, H_{\bar{S}}) \sim (F_{S'}, H_{\bar{S}'})} \frac{\Pr[F_S, H_{\bar{S}}]}{Z(F_S, H_{\bar{S}})} \sum_{(F_{S'}, H_{\bar{S}'}) \sim (F_{S''}, H_{\bar{S}''})} \frac{\Pr[F_{S'}, H_{\bar{S}'}]}{Z(F_{S'}, H_{\bar{S}'})} \cdot \Pr[F_{S''}, H_{\bar{S}''}] \cdot 2(m+2)\text{MPM}(F_{S''}, H_{\bar{S}''}) \\
&= 2(m+2)\text{MPM}(F_{S''}, H_{\bar{S}''}) \Pr[F_{S''}, H_{\bar{S}''}] \cdot \sum_{(F_S, H_{\bar{S}}) \sim (F_{S'}, H_{\bar{S}'})} \frac{\Pr[F_S, H_{\bar{S}}]}{Z(F_S, H_{\bar{S}})} \sum_{(F_{S'}, H_{\bar{S}'}) \sim (F_{S''}, H_{\bar{S}''})} \frac{\Pr[F_{S'}, H_{\bar{S}'}]}{Z(F_{S'}, H_{\bar{S}'})} \\
&\leq 2(m+2)\text{MPM}(F_{S''}, H_{\bar{S}''}) \Pr[F_{S''}, H_{\bar{S}''}] \cdot \sum_{(F_S, H_{\bar{S}}) \sim (F_{S'}, H_{\bar{S}'})} \sum_{(F_{S'}, H_{\bar{S}'}) \sim (F_{S''}, H_{\bar{S}''})} \frac{\Pr[F_{S'}, H_{\bar{S}'}]}{Z(F_{S'}, H_{\bar{S}'})} \\
&\leq 2(m+2)\text{MPM}(F_{S''}, H_{\bar{S}''}) \Pr[F_{S''}, H_{\bar{S}''}] \cdot m^2 \sum_{(F_{S'}, H_{\bar{S}'}) \sim (F_{S''}, H_{\bar{S}''})} \frac{\Pr[F_{S'}, H_{\bar{S}'}]}{Z(F_{S'}, H_{\bar{S}'})} \\
&\leq 2(m+2)\text{MPM}(F_{S''}, H_{\bar{S}''}) \Pr[F_{S''}, H_{\bar{S}''}] \cdot m^2 \sum_{v \in \bar{S}''} \frac{\Pr[v \notin H(v)]}{\Pr[v \in H(v)]} \\
&\leq \frac{\epsilon}{6} \text{MPM}(F_{S''}, H_{\bar{S}''}) \Pr[F_{S''}, H_{\bar{S}''}]
\end{aligned}$$

Note that for each  $(F_{S'}, H_{\bar{S}'})$ , there are at most  $m^2$  different  $(F_S, H_{\bar{S}})$  such that  $(F_S, H_{\bar{S}}) \sim (F_{S'}, H_{\bar{S}'})$ . So we have the second inequality. The third inequality can be seen by canceling out same multiplicative terms from the numerators and the denominators, as in Lemma 5.  $\square$

Therefore, we have obtained the main theorem in this section.

**Theorem 2** *Assuming there are even number of vertices in the stochastic graph, there exists an FPRAS for estimating the expected length of the minimum perfect matching.*

## 4 Minimum Cycle Covers

In this section, we consider the minimum length cycle cover problem. In the deterministic version of the cycle cover problem, we are asked to find a collection of vertex-disjoint cycles such that every vertex is in one cycle and the total length is minimized. Here we assume that every cycle contains at least two nodes. If a cycle contains exactly two nodes, the length of the cycle is two times the distance between these two nodes. The problem can be solved in polynomial time by reducing the problem to a minimum bipartite perfect matching problem<sup>1</sup> W.l.o.g., we assume no two edges in  $\mathcal{P} \times \mathcal{P}$  have the same length. For ease of exposition, we assume that for each point, there is only one node that may realize at this point. In principle, if more than one nodes may realize at the same point, we can create multiple copies of the point co-located at the same place, and impose a distinct infinitesimal distance between every pair of copies, to ensure no two edges have the same distance.

We need the notion of the nearest neighbor graph, denoted by  $\text{NN}$ . For an undirected graph, an edge  $e = (u, v)$  is in the nearest neighbor graph if  $u$  is the nearest neighbor of  $v$ , or vice versa. We also use  $\text{NN}$  to denote its length.  $\mathbb{E}[\text{NN}]$  can be computed exactly in polynomial time [23]. As a warmup, we first show that  $\mathbb{E}[\text{NN}]$  is a 2-approximation of  $\mathbb{E}[\text{CC}]$  in the following lemma.

---

<sup>1</sup>If we require each cycle consist at least three nodes, the problem is still poly-time solvable by a reduction to minimum perfect matching by Tutte [31]. Hartvigsen [21] obtained a polynomial time algorithm for minimum cycle cover with each cycle having at least 4 nodes Cornuéjols and Pulleyblank [13] have reported that Papadimitriou showed the NP-completeness of minimum cycle cover with each cycle having at least 6 nodes.

**Lemma 10**  $\mathbb{E}[\text{NN}] \leq \mathbb{E}[\text{CC}] \leq 2\mathbb{E}[\text{NN}]$ .

*Proof:* We show  $\text{NN} \leq \text{CC} \leq 2\text{NN}$  for each possible realization. We prove the first inequality. For each node  $u$ , there are two edges incident on  $u$ . Suppose they are  $e_{u1}$  and  $e_{u2}$ . We have  $\text{CC} = \frac{\sum_u (\text{d}(e_{u1}) + \text{d}(e_{u2}))}{2} \geq \text{NN}$ . The second inequality can be seen by doubling all edges in NN and the triangle inequality.  $\square$

We denote the longest edge in NN (and also its length) by  $L$ . Note that  $L$  is also a random variable. By the law of total expectation, we estimate  $\mathbb{E}[\text{CC}]$  based on the following formula:

$$\mathbb{E}[\text{CC}] = \sum_{e \in \mathcal{P} \times \mathcal{P}} \Pr[L = e] \cdot \mathbb{E}[\text{CC} \mid L = e]$$

It is obvious to see that  $\frac{\text{NN}}{n} \leq L \leq \text{NN}$ . Combined with Lemma 10, we have that

$$\text{d}(e) \leq \mathbb{E}[\text{CC} \mid L = e] \leq 2n\text{d}(e). \quad (1)$$

However, it is not clear to us how to estimate  $\Pr[L = e]$  and how to take samples conditioning on event  $L = e$  efficiently. To circumvent the difficulty, we consider some simpler events and condition  $L = e$  on those simpler events. Consider a particular edge  $e = (s, t) \in \mathcal{P} \times \mathcal{P}$ . Denote  $N_s(t)$  as the event that the nearest neighbor of  $s$  is  $t$  and  $N_t(s)$  as the event that the nearest neighbor of  $t$  is  $s$ . Let  $L_{st}$  be the event the longest edge  $L$  in NN is  $e(s, t)$ . Let  $A_s(t) = N_s(t) \wedge L_{st}$ . First we write

$$\begin{aligned} \mathbb{E}[\text{CC} \mid L = e] \cdot \Pr[L = e] &= \mathbb{E}[\text{CC} \mid A_s(t) \vee A_t(s)] \cdot \Pr[A_s(t) \vee A_t(s)] \\ &= \mathbb{E}[\text{CC} \mid A_s(t)] \cdot \Pr[A_s(t)] + \mathbb{E}[\text{CC} \mid A_t(s)] \cdot \Pr[A_t(s)] \\ &\quad - \mathbb{E}[\text{CC} \mid A_s(t) \wedge A_t(s)] \cdot \Pr[A_s(t) \wedge A_t(s)] \end{aligned}$$

Now, we show how to estimate  $\mathbb{E}[\text{CC} \mid A_s(t)] \cdot \Pr[A_s(t)]$  for each edge  $e(s, t)$ . The other two terms can be estimated in the same way. Also notice that the third term is less than the first and the second terms. Therefore, for any points  $s$  and  $t$ , we have the following fact which will be useful later:

$$\mathbb{E}[\text{CC}] \geq \mathbb{E}[\text{CC} \mid L = e] \cdot \Pr[L = e] \geq \mathbb{E}[\text{CC} \mid A_s(t)] \cdot \Pr[A_s(t)]. \quad (2)$$

Moreover, we have that

$$\mathbb{E}[\text{CC} \mid A_s(t)] \cdot \Pr[A_s(t)] = \mathbb{E}[\text{CC} \mid A_s(t)] \cdot \Pr[L_{st} \wedge N_s(t)] = \mathbb{E}[\text{CC} \mid A_s(t)] \cdot \Pr[L_{st} \mid N_s(t)] \cdot \Pr[N_s(t)]$$

Suppose  $v$  is the only point that may realize to  $s$  and  $u$  is the only point that may realize to  $t$ . We use  $B$  as a shorthand notation for  $B(s, \text{d}(s, t))$ . We first observe that  $\Pr[N_s(t)]$  can be computed exactly in poly-time as follows:

$$\Pr[N_s(t)] = p_{vs} \cdot p_{ut} \cdot \prod_{w \neq v, u} (1 - p_w(B))$$

Also note that we can take samples conditioning on the event  $N_s(t)$  (the corresponding probability distribution for node  $v$  is:  $\Pr[v \models r \mid N_s(t)] = \frac{p_{vr}}{1 - p_w(B)}$ ).

Estimating  $\mathbb{E}[\text{CC} \mid A_s(t)] \cdot \Pr[L_{st} \mid N_s(t)]$ : Next, we show how to estimate  $\mathbb{E}[\text{CC} \mid A_s(t)] \cdot \Pr[L_{st} \mid N_s(t)]$ . The high level idea is the following. We take samples conditioning on  $N_s(t)$ . If  $\Pr[L_{st} \mid N_s(t)]$  is large (i.e., at least  $1/\text{poly}(nm)$ ), we can get enough samples satisfying  $L_{st}$ , thus  $A_s(t)$ . Therefore, we can get  $(1 \pm \epsilon)$ -approximation for both  $\Pr[L_{st} \mid N_s(t)]$  and  $\mathbb{E}[\text{CC} \mid A_s(t)]$  in poly-time (we also use the fact that if  $A_s(t)$  is true,  $\text{CC}$  is at least  $\text{d}(s, t)$  and at most  $n\text{d}(s, t)$ ). However, if  $\Pr[L_{st} \mid N_s(t)]$  is small, it is not clear how to obtain a reasonable estimate of this value. In this case, we show the contribution of the term to

our final answer is extremely small and even an inaccurate estimation of the term will not affect our answer in any significant way with high probability.

Now, we elaborate the details. We iterate the following steps for  $N$  times ( $N = O(\frac{n^2 m^3}{\epsilon^3}(\ln n + \ln m))$  suffices).

- Suppose we are in the  $i$ th iteration. We take a sample  $G_i$  of the stochastic graph conditioning on the event  $N_s(t)$ . We compute the nearest neighbor graph  $NN(G_i)$  and the minimum length cycle cover  $CC(G_i)$  of  $G_i$ . If  $e(s, t)$  is the longest edge in  $NN(G_i)$ , let  $I_i = 1$ . Otherwise  $I_i = 0$ .

Our estimate of  $\mathbb{E}[CC \mid A_s(t)] \cdot \Pr[L_{st} \mid N_s(t)]$  is the following:

$$\left( \frac{\sum_{i=1}^N I_i \cdot CC(G_i)}{\sum_{i=1}^N I_i} \right) \left( \frac{\sum_{i=1}^N I_i}{N} \right) = \frac{\sum_{i=1}^N I_i \cdot CC(G_i)}{N}$$

It is easy to see the expectation of  $\frac{\sum_{i=1}^N I_i \cdot CC(G_i)}{N}$  is exactly  $\mathbb{E}[CC \mid A_s(t)] \cdot \Pr[L_{st} \mid N_s(t)]$ .

We distinguish the following two cases:

1.  $\Pr[L_{st} \mid N_s(t)] \geq \frac{\epsilon}{2nm^3}$ . By Lemma 1,  $\frac{\sum_{i=1}^N I_i}{N} \in (1 \pm \epsilon)\Pr[L_{st} \mid N_s(t)]$  with high probability. Moreover, we can get  $\sum_{i=1}^N I_i \geq \Omega\left(\frac{n}{\epsilon^2}(\ln n + \ln m)\right)$  with high probability. In this case, we have enough successful samples (samples with  $I_i = 1$ ) to guarantee that  $\frac{\sum_{i=1}^N I_i CC(G_i)}{\sum_{i=1}^N I_i}$  is a  $(1 \pm \epsilon)$ -approximation of  $\mathbb{E}[CC \mid A_s(t)]$  with high probability, again by Lemma 1. We note that under condition  $A_s(t)$ , we have a nice instance since  $CC$  is at least  $d(s, t)$  and at most  $nd(s, t)$ .
2.  $\Pr[L_{st} \mid N_s(t)] < \frac{\epsilon}{2nm^3}$ . We note that  $I_i = 0$  means that while  $N_s(t)$  happens, the longest edge  $L$  in  $NN$  is longer than  $e(s, t)$ . Suppose  $e(s', t')$  is the edge with the maximum  $\Pr[L_{s't'} \mid N_s(t)]$ . Since  $\Pr[L_{st} \mid N_s(t)] \leq \frac{\epsilon}{2nm^3}$ ,  $e(s', t')$  must be different from  $e(s, t)$  and  $\Pr[L_{s't'} \mid N_s(t)] \geq \frac{2nm^2}{\epsilon}\Pr[L_{st} \mid N_s(t)]$ . Hence, we have that

$$\begin{aligned} \mathbb{E}[CC \mid A_s(t)] \cdot \Pr[A_s(t)] &= \mathbb{E}[CC \mid A_s(t)] \cdot \Pr[L_{st} \mid N_s(t)] \cdot \Pr[N_s(t)] \\ &\leq n \cdot d(s, t) \cdot \frac{\epsilon}{2nm^2} \cdot \Pr[L_{s't'} \mid N_s(t)] \cdot \Pr[N_s(t)] \\ &\leq \frac{\epsilon}{2m^2} \cdot d(s', t') \cdot \Pr[L_{s't'} \mid N_s(t)] \cdot \Pr[N_s(t)] \\ &\leq \frac{\epsilon}{2m^2} \cdot \mathbb{E}[CC \mid A_{s'}(t')] \cdot \Pr[L_{s't'}] \\ &\leq \frac{\epsilon}{2m^2} \cdot \mathbb{E}[CC] \end{aligned}$$

The first and third inequalities are due to (1) and the fourth are due to (2). By Chernoff Bound, we have that

$$\Pr \left[ \frac{\sum_{i=1}^N I_i \cdot CC(G_i)}{N} \geq \frac{\epsilon}{m^2} \cdot \mathbb{E}[CC] \right] \leq \frac{e^{-n}}{m^2}$$

Then, with probability at least  $1 - \text{poly}(\frac{1}{n})$ , the contribution from all such edges is less than  $\epsilon \mathbb{E}[CC]$ .

In summary, we obtain the following theorem.

**Theorem 3** *There is an FPRAS for estimating the expected length of the minimum cycle cover in a stochastic graph.*

Finally, we remark that our algorithm also works in presence of both locational uncertainty and node uncertainty, i.e., the existence of each node is a Bernoulli random variable. It is not hard to extend our technique to handle the case where each cycle is required to contain at least three nodes. This is done by considering the longest edge in the 2NN graph (each vertex connects to the nearest and second nearest neighbors). The extension is fairly straightforward and we omit the details here.

## 5 Conclusion

We obtain FPRAS the problems of computing the expected lengths of the minimum spanning tree, the minimum perfect matching and the minimum cycle cover on a stochastic graph where the location of each node is a random point in a given metric space. Our results for the stochastic minimum perfect matching and the stochastic minimum cycle cover are the first known algorithms.

An interesting open problem is the problem of estimating the expected value of the minimum cost matching of a certain cardinality (instead of the perfect matching). It is not clear how to extend our technique to handle this problem. There are some other important combinatorial optimization problems that have not been studied in this model, such as the  $b$ -matching and the Euclidean  $k$ -median problem (the deterministic version admits a PTAS in Euclidean spaces [5, 25]).

It is also interesting to study problems for which the deterministic version is APX-hard. In such cases, it is not possible to obtain FPRAS and the best ratio we can hope for is the best approximation ratio we can obtain for the deterministic version of the problem.

## References

- [1] P. Afshani, P.K. Agarwal, L. Arge, K.G. Larsen, and J.M. Phillips. (approximate) uncertain skylines. In *Proceedings of the 14th International Conference on Database Theory*, pages 186–196. ACM, 2011.
- [2] P.K. Agarwal, S.W. Cheng, Y. Tao, and K. Yi. Indexing uncertain data. In *Proceedings of the twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 137–146. ACM, 2009.
- [3] P.K. Agarwal, A. Efrat, S. Sankararaman, and W. Zhang. Nearest-neighbor searching under uncertainty. In *Proceedings of the 31st symposium on Principles of Database Systems*, pages 225–236. ACM, 2012.
- [4] C. Alexopoulos and J.A. Jacobson. State space partition algorithms for stochastic systems with applications to minimum spanning trees. *Networks*, 35(2):118–138, 2000.
- [5] S. Arora, P. Raghavan, and S. Rao. Approximation schemes for euclidean  $k$ -medians and related problems. In *Proceedings of the thirtieth annual ACM symposium on Theory of Computing*, pages 106–113. ACM, 1998.
- [6] M.J. Atallah, Y. Qi, and H. Yuan. Asymptotically efficient algorithms for skyline probabilities of uncertain data. *ACM Trans. Datab. Syst.*, 32(2):12, 2011.
- [7] J. Beardwood, J. H. Halton, and J. M. Hammersley. The shortest path through many points. In *Proc. Cambridge Philos. Soc.*, pages 55:299–327, 1959.
- [8] M. W. Bern and D. Eppstein. Worst-case bounds for suadditive geometric graphs. In *Symposium on Computational Geometry*, pages 183–188, 1993.

- [9] D.J. Bertsimas and G. van Ryzin. An asymptotic determination of the minimum spanning tree and minimum matching constants in geometrical probability. *Operations Research Letters*, 9(4):223–231, 1990.
- [10] G. Beskales, M. Soliman, and I. Ilyas. Efficient search for the top-k probable nearest neighbors in uncertain databases. *VLDB*, 2008.
- [11] T.M. Chan. Well-separated pair decomposition in linear time? *Information Processing Letters*, 107(5):138–141, 2008.
- [12] R. Cheng, J. Chen, M. Mokbel, and C. Chow. Probabilistic verifiers: Evaluating constrained nearest-neighbor queries over uncertain data. In *ICDE*, 2008.
- [13] G. Cornuejols and W. Pulleyblank. A matching problem with side constraints. *Discrete Math.*, 29, 1980.
- [14] A. Czumaj, F. Ergün, L. Fortnow, A. Magen, I. Newman, R. Rubinfeld, and C. Sohler. Approximating the weight of the euclidean minimum spanning tree in sublinear time. *SIAM Journal on Computing*, 35(1):91–109, 2005.
- [15] A. Czumaj and C. Sohler. Estimating the weight of metric minimum spanning trees in sublinear-time. In *Annual ACM Symposium on Theory of Computing: Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, volume 13, pages 175–183. Citeseer, 2004.
- [16] Y. Emek, A. Korman, and Y. Shavitt. Approximating the statistics of various properties in randomly weighted graphs. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1455–1467. SIAM, 2011.
- [17] A.M. Frieze. On the value of a random minimum spanning tree problem. *Discrete Applied Mathematics*, 10(1):47–56, 1985.
- [18] P. Gupta and P.R. Kumar. Critical power for asymptotic connectivity. In *Proceedings of the 37th IEEE Conference on Decision and Control*, volume 1, pages 1106–1110. IEEE, 1998.
- [19] P. Gupta and P.R. Kumar. The capacity of wireless networks. *IEEE Transactions on Information Theory*, 46(2):388–404, 2000.
- [20] M. Haenggi, J.G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti. Stochastic geometry and random graphs for the analysis and design of wireless networks. *IEEE Journal on Selected Areas in Communications*, 27(7):1029–1046, 2009.
- [21] D. Hartvigsen. An extension of matching theory. phd thesis, carnegie-mellon university. 1984.
- [22] P. Kamousi, T. Chan, and S. Suri. Closest pair and the post office problem for stochastic points. *Algorithms and Data Structures Symposium*, pages 548–559, 2011.
- [23] P. Kamousi, T.M. Chan, and S. Suri. Stochastic minimum spanning trees in euclidean spaces. In *Proceedings of the 27th annual ACM symposium on Computational Geometry*, pages 65–74. ACM, 2011.
- [24] H. J. Karloff. How long can a euclidean traveling salesman tour be? In *J. Discrete Math*, page 2(1). SIAM, 1989.

- [25] S.G. Kolliopoulos and S. Rao. A nearly linear-time approximation scheme for the euclidean k-median problem. *SIAM Journal on Computing*, 37(3):757–782, 2007.
- [26] J. Li and A. Deshpande. Ranking continuous probabilistic datasets. *Proceedings of the VLDB Endowment*, 3(1-2):638–649, 2010.
- [27] M. Löffler. Data imprecision in computational geometry. 2009.
- [28] M. Löffler and J. Phillips. Shape fitting on point sets with probability distributions. *European Symposia on Algorithms*, pages 313–324, 2009.
- [29] T. L. Snyder and J. M. Steele. A priori bounds on the euclidean traveling salesman. In *J. Comput*, page 24(3). SIAM, 1995.
- [30] J.M. Steele. On frieze’s  $\zeta(3)$  limit for lengths of minimal spanning trees. *Discrete Applied Mathematics*, 18(1):99–103, 1987.
- [31] W. T. Tutte. A short proof of the factor theorem for finite graphs. *Canad. J. Math.*, 6, 1954.
- [32] H. Yıldız, L. Foschini, J. Hersherberger, and S. Suri. The union of probabilistic boxes: Maintaining the volume. *European Symposia on Algorithms*, pages 591–602, 2011.

## A Another FPRAS for MST

W.l.o.g., we assume that for each point, there is only one node that may realize to this point. Our algorithm is a slight generalization of the one proposed in [23]. Let  $\mathbb{E}[i]$  be the expected MST length conditioned on the event that all nodes  $\{s_1, \dots, s_n\}$  are realized to points in  $\{u_i, \dots, u_m\}$  (denote the event by  $\text{In}(i, m)$ ). Let  $\mathbb{E}'[i]$  be the expected MST length conditioned on the event that all nodes  $\{s_1, \dots, s_n\}$  are realized to  $\{u_i, \dots, u_m\}$  and at least one node is realized to  $u_i$ . We use  $s \models u$  to denote the event that node  $s$  is realized to point  $u$ . It is easy to see that

$$\mathbb{E}[i] = \mathbb{E}'[i] \Pr[\exists s, s \models u_i \mid \text{In}(i, m)] + \mathbb{E}[i + 1] \Pr[\neg \exists s, s \models u_i \mid \text{In}(i, m)]$$

For a particular point  $u_i$ , we reorder the points  $\{u_i, \dots, u_m\}$  as  $\{u_i = r_i, \dots, r_m\}$  in increasing order of distance from  $u_i$ . Let  $\mathbb{E}'[i, j]$  be the expected MST length for all nodes conditioned on the event that all nodes are realized to  $\{r_i, \dots, r_j\}$  (denoted as  $\text{In}'(i, j)$ ) and  $\exists s, s \models u_i$ . Let  $\mathbb{E}''[i, j]$  be the expected MST length for all nodes conditioned on the event  $\text{In}'(i, j) \wedge (\exists s, s \models u_i) \wedge (\exists s', s' \models r_j)$ . We can see that

$$\begin{aligned} \mathbb{E}'[i, j] &= \mathbb{E}''[i, j] \Pr[\exists s', s' \models r_j \mid \text{In}'(i, j), \exists s, s \models u_i] \\ &\quad + \mathbb{E}'[i, j - 1] \Pr[\neg \exists s', s' \models r_j \mid \text{In}'(i, j), \exists s, s \models u_i] \end{aligned}$$

It is not difficult to see the probability  $\Pr[\exists s', s' \models r_j \mid \text{In}'(i, j), \exists s, s \models u_i]$  can be computed in polynomial time. Here we use the assumption that for each point, only one node that may realize to it. Moreover, we can also take samples conditioning on event  $\text{In}'(i, j) \wedge (\exists s, s \models u_i) \wedge (\exists s', s' \models r_j)$ . Therefore  $\mathbb{E}''[i, j]$  can be approximated within a factor of  $(1 \pm \epsilon)$  using the Monte Carlo method in polynomial time since it is a nice instance. The number of samples needed can be bounded by a polynomial.

We can easily generalize the above algorithm to the case where  $\sum_{j=1}^m p_{ij} \leq 1$ , i.e., node  $i$  may not be present with some probability. Indeed, this can be done by generalizing the definition of  $\text{In}(i, j)$  (and similarly  $\text{In}'(i, j)$ ) to be the event that each nodes is either not present or realized to some node in  $\{r_i, \dots, r_j\}$ .